Chapter 10
Adapting the U.S. National Hydrography Dataset to Linked Open Data

Dalia E. Varanka, E. Lynn Usery and David M. Mattli

Abstract A controlled vocabulary for the National Hydrography Dataset (NHD) of the United States was developed as Linked Open Data (LOD). The vocabulary has two main parts: a glossary and a set of triples reflecting the NHD data model as it is organized in geographic information systems (GIS). The glossary consists of a feature type label and a comment consisting of a definition that is linked to a hydrographic feature type standard. The ontology of the data model consists of classes and properties that group and relate sets of individual features. The objective of the project is to draw on the glossary and the "triplified" data model to build formal semantics for a basic form of NHD as LOD. Modifications were made primarily to the specification of feature types for the data.

Keywords US hydrography dataset · Linked open data · Adapting

1 Introduction

Geospatial surface water data, a subset of the general field called hydrography, is a central theme of scientific, policy, and public interest. In the United States, the National Hydrography Dataset (NHD) is used for academic research, regulatory action by the Environmental Protection Agency, Bureau of Reclamation, and other agencies, and citizen-based projects such as local habitat restoration. Presently, most hydrographic data takes the form of a geographic information systems (GIS) database, an expensive technology that requires specialized training. Other guidelines that support understanding the data and decisions with regard to their use

D.E. Varanka (✉) · E. Lynn Usery · D.M. Mattli
U.S. Geological Survey, Rolla, Missouri, USA
e-mail: dvaranka@usgs.gov

E. Lynn Usery
e-mail: usery@usgs.gov

D.M. Mattli
e-mail: dmmattli@usgs.gov

C. Robbi Sluter et al. (eds.), *Cartography - Maps Connecting the World*,
Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-319-17738-0_10

take the form of natural language documentation, such as data standards or metadata files. Documentation is a labor-intensive development, in the sense that metadata and other context-building attachments may or may not be available; when they are, they present additional challenges for digital access.

Although GIS is a powerful tool for geospatial analysis, semantic technology capabilities suggest possibilities of improved aspects of data management and information query results. Applied ontology is regarded by many to offer intuitive data access and recognition (Mark et al. 2005). The graph model allows as many connections among objects or literals as needed, allowing for more direct access to a wider range of data relationships than are easily available in GIS. Unlike relational table databases, the relationships of the applied ontology supported by the triple data model can alter their schemas at run-time. SPARQL queries match triple-pattern resources individually and so are more directly processed and less complicated to form. Queries can draw from and construct complex classification systems and specific semantic conditions.

The objective of this study is to develop a controlled vocabulary reflecting the NHD of the United States that increases accessibility to data, and provides data flexibility and semantic specification. The approach is to publish a controlled vocabulary as Linked Open Data (LOD) freely available over the Internet. Vocabularies for LOD formalize semantic terms with logic specifications, enable machine processing, and include human readable annotation. The development of LOD based on a large-scale GIS dataset requires the integration of verbal documentation and tabular vector data for semantic formalization. Consequently, the design of the semantics takes what is often called a top-down approach that begins with general ontology concepts, or bottom-up approaches contributed from multiple individuals or observations. Further refinement is made through interactions with the data.

Broad principles exist that specify guidelines for LOD as a finished product (Berners-Lee 2006) or that exemplify LOD (Dbpedia 2012). The data integrate through the use of the Resource Development Framework (RDF) data model and related technology such as triplestores, SPARQL Protocol And RDF Query Language (SPARQL) endpoints, and Internet services. Network linkage of the data is facilitated by well-designed Universal Resource Identifiers (URIs). The data formalizations should stay as simple as possible to maximize reuse for multiple applications. The Ordnance Survey (OS) has made available LOD for topographic features (Ordnance Survey 2014; Hart and Dolbear 2013). Other LOD offer topographic feature vocabulary with URI, but specifications are limited to natural language terms (Smethurst et al. 2014).

Four characteristics of the triple data model take advantage of cognitive, technical, and geospatial aspects of spatial management and retrieval:

- Improved formal and informal semantic specification of knowledge
- Schema flexibility for database analysis and change
- Cognitively recognizable geospatial feature representation
- Rule and reasoning processing through inference engines

Structured Query Language (SQL) is the predominant data manipulation language for relational table data such as GIS. The range of query syntax is impressive, including sorting, restriction, selecting, joining, division, and Boolean and algebraic operations. Relates between tables can create queries that go beyond the table schema to enable spatial and nonspatial attribute data values. Spatial analysis can be applied visually with the graphic mapping medium, capturing analogue search results. Challenges for the relational table model, however, have long been recognized. These include the paucity of semantic detail to resemble cognitive recognition of real-world experiences and that the relational model lacks the efficient performance levels for geospatial processing (Worboys 1999). The relational table data model used by most GIS hinders the accuracy and precision of geospatial data. Data must conform to pre-determined attribute table designs that can't be easily modified for differing values. Relates between tables are limited to a subset of the possible attributes resulting in duplication of data in various tables. Relational table queries rely on Boolean operators that select columns as a whole for results. To return parts of a table (some of the column values), expressions must be combined or subqueries must be formed.

The triple data model aligns with cognitive spatial thinking in that each object is independently modeled based on a number of primitive elements of its criteria as a knowledge category. Dynamic binding operations, meaning multiple relations with other objects in the database, can be defined for a single entity, allowing that entity to assume multiple qualities and roles. When many relations are possible as criteria of class definitions, axioms for their logical association are required. Although in many domains of expertise associations are debatable or subjective, spatial queries offer the advantage of referring to real-world entities that occupy time and space. Geospatial rules and definitions have been developed in the body of literature of geography and geographic information science, but they are poorly represented as yet in formal semantics. The simplicity of the triple data model allows not only the representation of complex semantics but also improved integration of graphs and their subsets between databases over networks built by multiple and diverse populations.

The design of the NHD LOD vocabulary combines the perspectives articulated in LOD research but identifies varying semantic levels. The concept of semantic levels is consistent with earlier theoretical research, although the levels themselves may vary. The design adapted for this project maintains the ability to use original data instances, but it modifies classes and properties to create categories that serve as objects of discourse at different stages of data modeling. As a result, the approach in this paper describes NHD at different semantic perspectives. These include the following:

- Controlled vocabulary: natural language definitions that were developed for general hydrographic features collected for and depicted on topographic maps
- Triplification: the conversion of the GIS attribute table schemas to RDF
- Formal semantics: manual mediation of the converted data to a linked data design in RDF

- LOD: integration between the NHD vocabulary and geospatial feature data instances
- Subject domain: the potential integration of the NHD data vocabulary with a general ontology for surface water

Controlled vocabularies have a history preceding LOD in data management practices. Controlled vocabularies can be defined as the representation of a limited set of terms and various relations among them that are maintained to achieve consistency in use. They may include a definition (a gloss); related terms, as in a thesaurus; or their formal semantic specifications, such as logic statements within an ontology. The design of controlled vocabularies supports the disambiguation and interchange of data. The National Information Standards Organization (NISO) recommends preferred techniques and procedures for displaying terms in a controlled vocabulary (NISO 2005). The analysis of vocabularies in this standard is applicable to, but wider than, ontology. The different structures described in the standard, such as indexes, facets, warrants, taxonomy, thesauri, and other relations of terms-to-context, clearly help organize ontology. Guidelines for the use of controlled vocabulary with metadata schemas are offered. Approaches to disambiguation are explained, addressing different parts of speech, punctuation, and ontological relationships.

The LOD files are relatively simple and limited to the basic concepts such as classes and subclasses; and annotation, object, and datatype properties. They are connected by unique URIs and also joined within sets by type. The LOD vocabulary provides a generalized level of category criteria within which instances of data can be included. The NHD feature data instances, however, are accessed separately, downloaded from a viewer for The National Map of the U.S. Geological Survey (USGS); the project focuses on schema-level connections, thus avoiding the storage of converted data (Bizer et al. 2009). The vocabularies are also data, as they are instances of categories or members of a classification system or schema.

A particular focus of the analysis is explained with regard to the NHD FCode, a table of codes that indicate feature type categories. The FCode table example used in this chapter demonstrates that attributes for Pipeline features are selectively and repetitively arranged in cross-reference to each other to create subclasses. The example was reorganized to semantically specified feature type classes. The FCode design was initially constrained by the available technology, but their reorganization as LOD demonstrates the flexibility now available with semantic technology.

Specifying the semantic resolution of the LOD geospatial features is guided by the design of an ontology pattern (OP) informed by geographic information science theory. An OP is an abstract model of essential types and properties of real world features that are observed repeatedly from instance to instance. An OP is a graph-based data model that formalizes logical relations between elements that include the necessary and sufficient condition for statements of fact. The simple and universal conditions that form the minimum semantics of a concept can thus be reused to link to the same or very similar core concepts in data applications (Gangemi and Presutti 2010). Core concepts typically consist of simple primitives,

so that complex terms can be developed or arranged from the primitives, yet the use of OP in complex applications assist data integration because these commonly shared core conceptualizations are linked as equivalent to each other. The OP method is conducive to basic specification of commonly used concepts that can be reused and customized. The pattern could be considered to be an information object in that it is location independent and represents information independent of a specific copy. Paraphrased, a geographic information construct is a relational concept, not directly representational of real entities, but presupposes an intelligent recipient/decoder as well as a source (Couclelis 2010, 1786). With the application of the OP, its conceptual design will be loosened to address discontinuities and to address practical issues.

The sections of this chapter are organized as follows: Details of the approach are in the next section, including the development of namespaces, the two vocabularies used to develop the NHD feature semantics design. Conclusions are drawn at the end.

2 Approach

The NHD LOD semantic development began with a natural language glossary. The glossary highlights the feature-based approach that is typical of topographic data. Following this, an ontology of the NHD data model as it has been developed in GIS was manually modified to align to a surface water ontology, while maintaining links to data instances. The objective is to link NHD data converted from GIS to a range of surface water applications.

The natural language definitions forming the glossary use RDF annotation properties. Annotation properties are useful for humans but cumbersome in the database. The semantics of geospatial feature concepts were selected to use as a model. Specific modifications were made to the model after content analysis of supporting documentation. After some modification, the vocabulary was tested by querying the NHD vocabulary with triple data converted directly from GIS.

Data queries are made more complex by association with geospatial feature geometry objects. Because tables store lines, and by extension, polygons, as coordinate pairs in a single line, a geometry object is degraded to a series of segments from node to node with no regard to the beginning and end of a geometry object. SQL returns query results as lists of rows that match a specific condition, outside of the geometry of the feature object itself. As a result, features are often classified by the geometry type, such as points or areas that can be supported in a single GIS layer consisting of a table instead of as feature types known by natural language terms, such as river, road, or marsh. In contrast to this processing sequence, SPARQL matches triple resources of each tuple, enabling combinations of features with geometry types. Being set based, classes can maneuver groups of features without regard to their topological structures. Relative spatial relations can

be formalized using descriptive logic or first-order logic, and thus advance spatial queries in a way unattained by GIS.

Although the term *feature*, meaning geospatial feature type, is widely used in the geospatial data community, its semantic specifications can be inconsistent, even within the same data user community. For example, one potential application for a vocabulary of geospatial feature types, from here forward called *features*, is for the feature type specification common to gazetteers. The International Organization for Standardization (ISO) offers spatial referencing specifications for gazetteers in which the element Feature has a direct relation to SpatialReference (ISO 19107:2003). Compliance with the ISO standard, however, initially seems to conflict with the GeoSPARQL standard ontology developed by the Open Geospatial Consortium (OGC) for geospatial data queries (Perry and Herring 2012). The GeoSPARQL standard ontology has a relation between the elements Feature and Geometry. Geometry is a subclass of SpatialReference in the ISO standard. Although such inconsistencies can be resolved by specifying Web Ontology Language (OWL) axioms and applying inference, the semantics of the term for purposes of this study will be clarified (W3C OWL Working Group 2012).

The semantics of Feature as a concept have been explored in the literature of geographic information science. Usery (2014) offered a history of the concept through multiple phases of specifications. Landmark literature establishes Feature as a triad of Theme, Space, and Time, equivalent to a geographical fact (Berry 1964). To specify a feature type class, elements of Attribute and Spatial Relation are added, as shown in Fig. 1.

The feature type class concept served as a model for selecting appropriate RDF properties and object classes, or datatype literals, although they were labeled in a way that is different from the feature concept literature.

Content analysis is the broad term for the analysis of texts. Analysis can take the various forms, although the study of texts for their meaning is a primary objective.

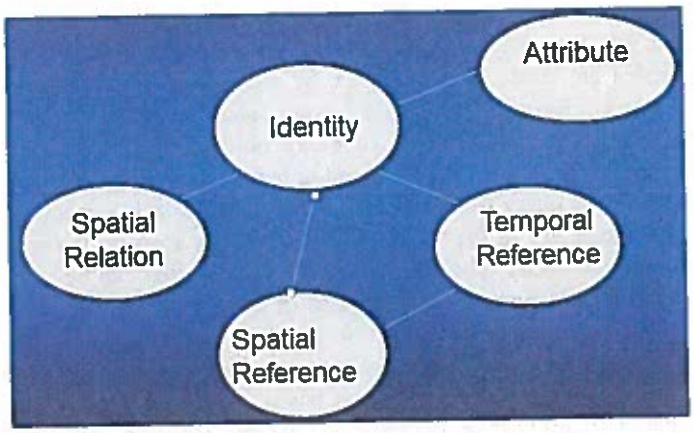

Fig. 1 Elements of geospatial feature type class concept

For the design of an NHD ontology, the NHD Users Guide (USGS 2014a, b, c) and the standard for the related Watershed Boundary Dataset (WBD) (USGS 2013) were primary sources. Comparison of text sources was sometimes made to pre-staged datasets available online. USGS specialists answered questions and provided guidance via email.

Content analysis of NHD/WBD sources consisted almost entirely of analyzing individual statements for specific semantic meaning. The triple data model is essentially the conversion of a simple, three-word sentence: subject, predicate, object. Although triples are simplified statements that do not resemble natural language, language can be formalized to triples using the RDF vocabulary. Graphic or tabular examples in the documents provided additional clarification of language semantics. All content analysis was manually accomplished.

The basis for designing and using LOD is effective URIs. The URI domain name for all three parts of the vocabulary is http://cegis.usgs.gov/. Two types of entities follow the domain name. The more common approach is a term indicating a resource for a broad subject domain, such as "Surface Water" or "Terrain," and then a dataset, such as "NHD." The name of a feature type, such as "Stream," follows the resource term or product name either as part of the path or as the data instance identifier, appearing last. A second option that is less often used is that the name of a major USGS product, such as the NHD, Geographical Names Information System (GNIS), or a set of standards follows the domain name. Such URIs are sometimes required for data handling uses, such as "gloss" for a feature type definition.

3 NHD LOD

3.1 Namespaces

Recommended URIs for LOD are different for representations of the real-world resource and for documents describing the resource and data. For example, a Hyper Text Markup Language (HTML) document would be available for human readability of the vocabulary contents, an RDF for machine readability of the contents, and a URI describing resources would be used for semantic negotiation and resolution. These last URIs would lack any extension at their end. The following example depicts the relation between the concept document and the RDF and HTML documents.

http://cegis.usgs.gov/surfacewater/Lake The resource URI
http://cegis.usgs.gov/surfacewater/Lake.rdf The RDF document URI
http://cegis.usgs.gov/surfacewater/Lake.html The HTML document URI

These URIs appear in the header of the RDF file that includes a list of prefixes to shorten the URI character string when in use. The combination of the prefix and the

name of a specific term enables a qualified names (qname) for the namespace. In the following example, the longer URI is substituted by the prefix SW in the namespace for Lake:

@prefix SW: http://cegis.usgs.gov/SurfaceWater
SW:Lake.

3.2 Glossary

The glosses, or natural language definitions, are only humanly readable. Taking the form of comments, these are stored in a separate file from the data because they technically consist of annotation properties that add unnecessary volume to data files that are meant to be processed primarily by machines. The glosses were initially defined in data standards and these standards are references for each geospatial feature type as a class. A separate URI was determined for standards developed within the USGS National Map enterprise.

@prefix NM: <http://nationalmap.gov/standards/>.
<SW:Stream>rdfs:isDefinedBy <NM:nhdstds.html>.

The informal semantics of the resource are formalized by applying specific logical relations called properties. Wherever possible, existing vocabularies were reused for feature classes and properties.

A controlled vocabulary was developed to support natural-language feature term glosses. Based on the same object class Feature as in the NHD data vocabulary, three basic properties are included: rdfs:label, rdfs:comment (the definition), and rdfs:isDefinedBy. The feature type names and definitions are taken from the USGS mapping and digital data standards as members (USGS 1996). The domain of the rdfs:label and rdfs:comment properties is the feature type specified in NHD; it is this class of objects that will be linked. The label and comment must be linked to each other. Though the property rdfs:isDefinedBy applies to the label and comment, its domain class is also the feature. The Feature class may have more than one label and/or gloss, in which case label and comment would be blank nodes with the standard defining the term as the object of rdfs:isDefinedBy. The vocabulary name for a specific resource includes Dublin Core Metadata Initiative (DCMI) dcterms: title and dcterms:description.

3.3 GIS NHD

The NHD is an extensive dataset supporting a rich collection of terms. The LOD was created from the GIS by using a conversion program, content analysis, and ontology.

3.4 Automatic Conversion

The creation of the ontology began with NHD data conversion. The custom conversion program creates triples formed directly from the GIS attribute table by using subjects from rows, properties from column headings, and objects from cell values contained in NHD (Mattli 2013). For example, the schema listing the column names and types for NHDFlowline is depicted in Fig. 2. When serialized as triples, the subjects whose name appears in the Field Name column, such as "FDate" or "Resolution" would have the property "Data Type" and object value stored in the cell of the Data Type column, such as "Date/Time" or "Number."

The conversion program accepts a Personal Geodatabase (GDB) format (.mdb) file downloaded from The National Map as input. The GDB files are based on Microsoft Access tables with additional formatting for geospatial information. NHD tables are based on their assigned feature types which are organized by geometry, points, lines (arcs), and polygons. Output triples of The National Map data are formatted in the schema of RDF and use additional resources from other vocabularies of semantic technology terms. URIs are assigned to each resource and can be found in the header of the RDF document.

In the conversion program, a template is created for each targeted table of features creating classes of individual members. The relational data model of NHD stores segments of the spatial geometry of features as unique rows in a database table. For example, Fig. 3 depicts three rows having the name 'Joachim Creek.' These rows are three segments of the same river; they do not represent three rivers with the same name.

Fig. 2 NHDFlowline schema

Field Name	Data Type
OBJECTID	AutoNumber
Shape	OLE Object
ComID	Number
Permanent_Identifier	Text
FDate	Date/Time
Resolution	Number
GNIS_ID	Text
GNIS_Name	Text
LengthKM	Number
ReachCode	Text
FlowDir	Number
WBAreaComID	Number
WBArea_Permanent_Identifie	Text
FType	Number
FCode	Number
Shape_Length	Number
Enabled	Number

NHDFlowline

OBJECTID *	Shape *	ComID *	FDate	Resolution	GNIS_ID	GNIS_Name	LengthKM	ReachCode *
10	Polyline ZM	3626819	7/28/1999	Medium	<Null>	<Null>	1.793	07140101001293
11	Polyline ZM	3627019	7/28/1999	Medium	<Null>	<Null>	1.542	07140101001681
12	Polyline ZM	3627063	7/28/1999	Medium	<Null>	<Null>	1.544	07140101001640
13	Polyline ZM	3627373	7/28/1999	Medium	00756230	Joachim Creek	0.485	07140101000124
14	Polyline ZM	3627523	7/28/1999	Medium	00756230	Joachim Creek	0.675	07140101000135
15	Polyline ZM	3627737	7/28/1999	Medium	00756230	Joachim Creek	0.282	07140101000148
16	Polyline ZM	3627775	7/28/1999	Medium	<Null>	<Null>	0.659	07140101001049
17	Polyline ZM	3627911	7/28/1999	Medium	00756537	Platte Creek	0.324	07140101000185

Fig. 3 Features as defined in GIS are segments of a feature object

Each row representing a feature segment is converted to RDF. To assign a globally unique identifier to each row, the template first specifies a column to use as the unique part of the URI. For NHDFlowline, the "Permanent_Identifier" field is used. The conversion program combines this field with a partial (baseline) URI (http://cegis.usgs.gov/rdf/nhd/) to generate the final feature URI. For example, the "Permanent_Identifier" "102209808" becomes "http://cegis.usgs.gov/rdf/nhd/Features/102209808". For the remaining columns of the row, the triples are generated using this pattern: <Feature URI> <Column name> <Column value>. Column names in the triple property position are assigned URIs and a datatype is specified for the value, or if the value is an element of a structured vocabulary such as the GNIS, the value is assigned a URI as well.

Many column headings were converted to properties meaning new classes were created. Many 'attributes are applied to instances, especially in the gazetteer ontology, so the creation of a class is required to define them as concepts.

All of the attributes of each table were captured as object or datatype properties within a taxonomic class grouping for properties, with a few exceptions such as FCode and event type which are better represented as classes. A sample of converted data, showing the vocabularies applied in the URI and three hydrologic features with their associated hydrologic unit (HU) and geometries is copied below.

```
@prefix geo: <http://www.opengis.net/def/geosparql/>.
@prefix gnis: <http://cegis.usgs.gov/rdf/gnis/>.
@prefix hu: <http://cegis.usgs.gov/rdf/huc/>.
@prefix nhd: <http://cegis.usgs.gov/rdf/nhd/>.
@prefix nhdf: <http://cegis.usgs.gov/rdf/nhd/Features/>.
@prefix nhdg: <http://cegis.usgs.gov/rdf/nhd/Geometries/>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
<http://cegis.usgs.gov/rdf/hucf/0202> a hu:WBD_HU4;

    hu:hu4Name "Upper Hudson";
    hu:shapeArea 3.0;
    hu:shapeLength 18.0;
    geo:hasGeometry <http://cegis.usgs.gov/rdf/hucg/0202>.
```

<http://cegis.usgs.gov/rdf/hucf/020200> a hu:WBD_HU6;

hu:hu6Name "Upper Hudson";
hu:shapeArea 3.0;
hu:shapeLength 18.0;
geo:hasGeometry <http://cegis.usgs.gov/rdf/hucg/020200>.

<http://cegis.usgs.gov/rdf/hucf/02020004> a hu:WBD_HU8;

hu:hu8Name "Mohawk";
hu:shapeArea 0.0;
hu:shapeLength 9.0;
geo:hasGeometry <http://cegis.usgs.gov/rdf/hucg/02020004>.

No URIs were created for table names because NHD, like other GIS attribute tables, are organized by geometric classes. Features are constrained in the table by geometry classes, not as object types, as they are in the ontology. In addition to feature instances created by the conversion program, tables were converted to domain and range classes rather than geometry features. Statements of domain and range were manually added to data subjects and properties as part of the ontology. This allows all instances (rows) that share the same generated attribute values to be the restricted set (domain class) that the property can draw upon as the subject. For example, the NHD table called NHDVerticalRelationship has three column headings that were converted to properties to connect subjects (Permanent_Identifier) to the possible or allowed object values. One of those attributes that was converted to a property is Below_Permanent_Identifier. By establishing NHDVerticalRelationship as the domain class for Below_Permanent_Identifier, only members of NHDVertical_Relationship are useable subjects for Below_Permanent_Identifier property (Fig. 4).

The final set of classes and properties is listed in Fig. 5.

The general method was to refine the automatically converted data by identifying the rules expressed in the texts and representing them as restrictions on the properties between subject and object classes or literal instances. These rules that are otherwise embedded in documents were drawn from the NHD Users Guide and the WBD standard. In this way, the information that must otherwise be cognitively connected between documentation and databases by users was formalized as part of the data model so that a computer could automatically draw relationships.

3.5 *Linking NHD Feature Glosses to Feature Type Resources*

To link the data, a single triple was designed that makes the feature type class of the glossary equivalent to the feature type class of the dataset.

sw:nhd owl:equivalentClass sw:glossary

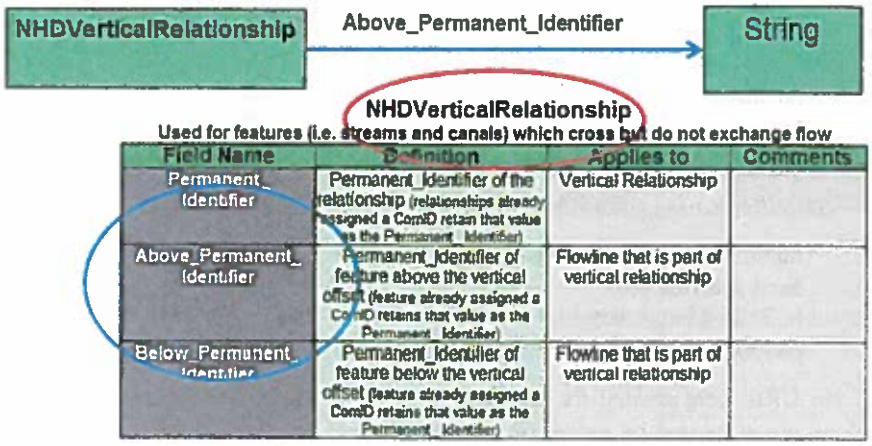

Fig. 4 In the case of NHDVerticalRelationship table, all instances (rows) of this class (table) can have a datatype property (column) of abovePermanentIdentifier to store a specific value. The domain (allowable subjects) of abovePermanentIdentifier are all the instances (rows) in the NHDVertical Relationship table having a (datatype string) value from the Above_Permanent_ Identifier column

Fig. 5 Classes and properties of GIS NHD

The two feature types are not interchangeable because one is a subclass of cegis.usgs.gov/surfacewater/nhd and the other is a subclass of cegis.usgs.gov/surfacewater/glossary. The feature types specified are matched through inference.

A second triple was necessary to link the NHD classes to instance members of the same, appropriate class. Using the prefix cegis for the namespace http://cegis.usgs.gov/rdf/, the triple to link classes to instances is:

sw:nhd owl:equivalentTo cegis:nhd.

3.6 *Feature Ontology Pattern*

In the GIS NHD, features are classified as an NHD FCode, an abbreviation meaning feature type code. The NHD FCode is a 5 digit numerical identifier for a feature type where the first three digits determined the type and the last two digits designate special attributes associated with it. For example, the FCode 42801 represents a pipeline indicated with 428 the 01 designating an aqueduct with a relationship of at or near the earth surface. FCodes are actually systematic combinations of a limited number of features types, attributes, and spatial relationships. A sample from the NHD FCode table is shown in Table 1.

FCodes are an example of a lookup table with limitations due to the design as a set matrix of attributes. Several aspects of such a table are problematic. Feature

Table 1 NHD FCode table for the pipeline feature type

Feature type	FCode	Description
PIPELINE	42800	Feature type only: no attributes
PIPELINE	42801	Pipeline type: aqueduct; relationship to surface: at or near
PIPELINE	42802	Pipeline type: aqueduct; relationship to surface: elevated
PIPELINE	42803	Pipeline type: aqueduct; relationship to surface: underground
PIPELINE	42804	Pipeline type: aqueduct; relationship to surface: underwater
PIPELINE	42805	Pipeline type: general; relationship to surface: at or near
PIPELINE	42806	Pipeline type: general; relationship to surface: elevated
PIPELINE	42807	Pipeline type: general; relationship to surface: underground
PIPELINE	42808	Pipeline type: general; relationship to surface: underwater
PIPELINE	42809	Pipeline type: penstock; relationship to surface: at or near
PIPELINE	42810	Pipeline type: penstock; relationship to surface: elevated
PIPELINE	42811	Pipeline type: penstock; relationship to surface: underground
PIPELINE	42812	Pipeline type: penstock; relationship to surface: underwater
PIPELINE	42813	Pipeline type: siphon; relationship to surface: unspecified
PIPELINE	42814	Pipeline type: general
PIPELINE	42815	Pipeline type: penstock
PIPELINE	42816	Pipeline type: aqueduct

types can be shared between classes in contradiction to principles of semantic categorization. For example, Aqueduct is categorized as either Canal or Pipeline, with a different feature code assigned for each, but this distinction contradicts the proper semantics of an aqueduct as either a watercourse system or a bridge. Types and attributes are duplicated across codes and attributes buried in the code are not available for specific information retrieval, only as a FCode group. A user querying the data cannot know the meaning of an FCode without manually looking it up in an online table.

The design of FCode tables was a result of technical constraints involving USGS data from the transition from analog to digital technologies. The codes were assigned to field-verified features in the Digital Line Graph vector datasets in the 1970s. DLG structured major and minor codes reflecting combinations of feature types and attributes. Because Fortran code as a mathematical language had limitations on character and world processing, text labels were specified in documentation. This technical arrangement for semantic information persists in the NHD. Semantic properties allow what NHD does not, as shown below.

The rdfs:subclass property, applied as part of the taxonomic hierarchy, allows for more diverse feature type and query connections than a table field for "Type" in GIS. "Types" derived from the GIS FCodes were made subclasses of appropriate feature classes in an ontology. Subclasses were defined as a specific type of a parent class, maintaining some qualities of the parent class but having additional specific qualities to differentiate it from its sibling classes. All of the attributes of each FType were captured as object or datatype properties available to any taxonomic class. For example, the subgraph for Pipeline FCodes depicted in Fig. 6 has asserted triples for the typeOf and relationshipToSurface properties for Pipeline. The typeOf relation is already asserted in a taxonomy (tree model), but the relationshipToSurface property allows the direct representation of Pipeline relation to the Earth's surface through the addition of a network model property, RelationshipToSurface, between these classes.

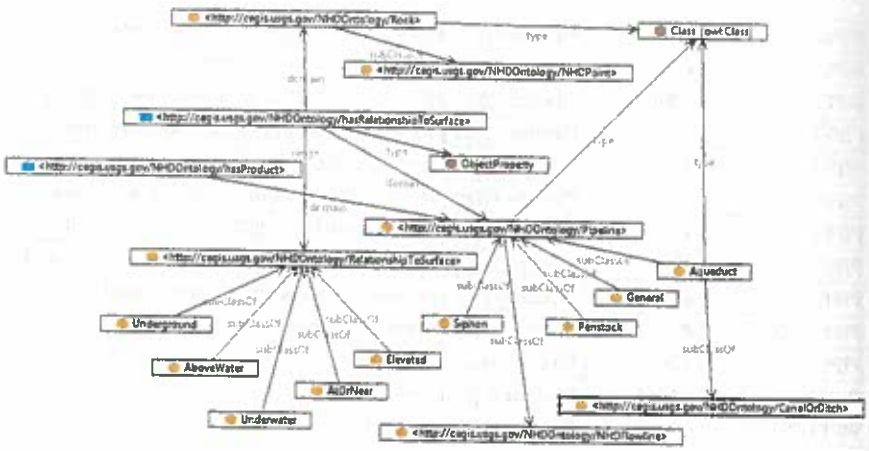

Fig. 6 Pipeline and relationship to surface subclasses modeled as a graph

As a result, triples that are equivalent to the natural language statement 'Members of the class Pipeline have a relation to Earth's surface and that relation is one of the set of RelationToSurface' is created. Instead of attributes being repeated, each class is entered into the graph triplestore only once. To demonstrate, The Aqueduct subclass is also a subclass of CanalOrDitch, though it is modeled only once instead of multiple times as in the FCode table. The Relationship to Surface class displays an additional relation to Rock. In addition to FCode semantics, the classes of the graph are also specified as instances of an ontology Class, as NHD geometry types, and properties can be specifieded as an object or datatype, as shown for the property hasRelationshipToSurface.

Ontologies offer improved semantic specification over GIS through the design and application of defined classes. Ontologies have two main types of classes: primitive classes that consist mainly of natural language terms within a hierarchy, and defined classes, specified as statements of criteria combinations. Defined classes take any number of properties to define the criteria that build a Feature concept. The following specifies a basic list of properties for a feature:

- Identity: rdf:typeOf
- Attribute: owl:equivalentTo
- Spatial reference: geo:hasGeometry
- Temporal reference: dc:date
- Spatial relations: the GeoSPARQL vocabulary

For all feature classes in the NHD ontology, GIS NHD terms with "type" as an attribute name, such as Pipeline Type, were classified in the taxonomy according to their possible ranges. Some FCode attributes were reordered as subclasses to other parent classes according to ontological distinctions. For example, many feature attributes from the FCode table took the form of spatial relations or qualities, rather than objects. Attribute terms were sometimes used as properties, not classes. For example, a column heading for construction material, although an object, is used only as part of the defined criteria of a feature. Properties of other vocabularies act to link the NHD to broadly used data. For example, a relevant property for the NHD that is not available in GIS is the dc term "hasPart" and "isPartOf:"

@prefix dcterms: <http://dublincore.org/documents/2012/06/14/dcmi-terms/>
<SW:Stream> dcterm:hasPart <SW:Streambank>.

FCodes were kept only as a resource to offer the option of linking legacy data and as a model for triples whose related semantics are to be formalized. FCodes themselves were eliminated because features can now have any number or type of properties.

3.7 *Linking Instance Data*

The NHD data for the vocabulary are converted on a case-by-case basis using a program available over the Internet. A user would use The National Map viewer to

select data, then convert it using a USGS program designed for that purpose. The namespace design described above can be used to connect concepts to data by using fragment identifiers. The instance identifier can be added to the end of the URI for the RDF vocabulary document without the fragment identifier, or with a # before the term for the instance identifier. The following two examples illustrate the two methods:

http://cegis.usgs.gov/surfacewater/stream/bigpiney
http://cegis.usgs.gov/surfacewater/stream#bigpiney

3.8 *Linking NHD Data to Other Vocabularies*

Much of the information that is unspecified in an OP is needed for practical applications. Related information to the OP is provided through important linkages to widely used ontology modules, such as the W3C PROV ontology for data provenance or the OGC Observations and Measurements (O&M) ontology (Lebo et al. 2013; Cox 2010). Terms unique to the NHD LOD are mapped to RDFS and OWL parent classes:

- Unique classes are mapped to rdfs:subclassOf
- Unique properties are mapped to rdfs:subPropertyOf
- Classes of two vocabularies are linked using owl:equivalentClass
- Equivalent properties of two vocabularies are linked using owl:equivalent Property

An incoming link pattern is found when the subject of a triple is not the object of any other triple. The object of the incoming link is the main concept or term being semantically described in the vocabulary. Triples that describe related resources are optional links; such triples expand the vocabulary term beyond the basics such as rdf:type, rdfs:label, or rdfs:isDefinedBy. Other vocabularies that are a good match to the LOD, having similar semantics, extensive coverage, and are widely used and maintained include geonames.org, a database and ontology about placenames, and LinkedGeoData (Geonames.org. 2014; Stadler et al. 2012).

4 Conclusions

The NHD is widely accessed by institutions and the general public to serve specific application needs, but users must be technically proficient in GIS and must learn a wide range of NHD-specific codes, keywords, and concepts. Impediments to data access were reduced by modeling the NHD as LOD formatted as RDF. One key example of data simplification and reuse is reorganizing NHD FCodes as an OP. This allows any number of optional properties instead of pre-defined statements

with inflexible constraints; any user is able to determine the feature type based on its contextual associations. That information can be returned automatically from a greater range of variables when querying the database. Linking the FCode that is stored as a value for each row of the data table to a natural language feature type term is an example of legacy technological constraints eliminated by semantic technology. The initial design of the NHD vocabulary is promising but requires further evaluation.

References

Berners-Lee T (2006) Linked data: W3C. http://www.w3.org/DesignIssues/LinkedData.html. Accessed 23 Sept 2014

Berry BJL (1964) Approaches to spatial analysis: a regional synthesis. Ann Assoc Am Geogr 54:2–11

Bizer C, Heath T, Berners-Lee T (2009) Linked data—the story so far. Int J Semant Web Inf Syst 5 (3):1–22

Couclelis H (2010) Ontologies of geographic information. Int J Geogr Inf Sci 24(12):1785–1809

Cox SJD (2010) Observations and measurements—XML implementation v2.0: OGC Implementation Standard 10-025. http://portal.opengeospatial.org/files/41510 Accessed 12 Nov 2014

Dbpedia (2012) Dbpedia. Universität Leipzig, University of Mannheim, and OpenLink Software. http://dbpedia.org/About. Accessed 14 Nov 2014

Gangemi A, Presutti V (2010) Towards a pattern science for the semantic web. Semant Web 1 (1–2):61–68

Geonames.org. (2014) GeoNames. http://www.geonames.org/. Accessed 12 Nov 2014

Hart G, Dolbear C (2013) Linked data: a geogrpahic perspective. CRC Press, Boca Raton

ISO 19107-2003, Geographic information-spatial schema. International organization for standardization (ISO). http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=26012

Lebo T, Sahoo S, McGuinness D (2013) PROV-O: the PROV ontology. http://www.w3.org/TR/prov-o/. Accessed 21 July 2014

Mark D, Smith B, Egenhofer MJ, Hirtle SC (2005) Ontological foundations for geographic information science. In: Robert b, McMaster, Usery EL (eds) A research agenda for geographic information science. CRC Press, Boca Raton

Mattli D (2013) NationalMap2rdf-new.py. In: Computer program: U.S. geological survey. http://cegis.usgs.gov/ontology.html. Accessed 29 May 2014

NISO (2005) ANSI/NISO Z39.19—Guidelines for the construction, format, and management of monolingual controlled vocabularies. http://www.niso.org/standards/resources/Z39-19.html

Ordnance Survey (2014) Ordnance survey linked data platform. http://data.ordnancesurvey.co.uk/

Perry M, Herring J (2012) OGC GeoSPARQL—a geographic query language for RDF data. Open geospatial consortium project document OGC 11-052r4, v. 1.0

Smethurst M, Styles R, Scott T (2014) The places ontology. http://vocab.org/places/schema.html

Stadler C, Lehmann J, Höffner K, Auer S (2012) LinkedGeoData: a core for a web of spatial open data. Semant Web 3(4):333–354. http://iospress.metapress.com/content/141w054666871326/

US Geological Survey (1996) Standards for 1:24,000-scale digital line graph and quadrangle maps: national mapping program technical instructions. http://nationalmap.gov/standards/dlgstds.html. Accessed 27 Feb 2014

US Geological Survey (2013) Ontology for the national map. http://cegis.usgs.gov/ontology.html

US Geological Survey (2014a) Hydrography; national hydrography dataset, watershed boundary dataset: U.S. geological survey. http://nhd.usgs.gov/. Accessed 18 Dec 2013
US Geological Survey (2014b) The national map: U.S. geological survey. http://nationalmap.gov/. Accessed 18 Dec 2013
US Geological Survey (2014c) NHD user guide. http://nhd.usgs.gov/userguide.html. Accessed 27 Feb 2014
Usery EL (2014) Spatial feature classes: encyclopedia of geography. Wiley, New York (In press)
W3C OWL Working Group (2012) OWL 2 web ontology language document overview, 2nd edn. W3C, Massachusetts. http://www.w3.org/TR/owl2-overview/. Accessed 14 Nov 2014
Worboys MF (1999) Relational databases and beyond. In: Longley PA, Goodchild MF, Maguire DJ, Rhind DW (eds) Geographical information systems, volume 1 principles and technical issues, 2nd edn. Wiley, New York